METHOD, SYSTEM, AND COMPUTER PROGRAM PRODUCT FOR REPRESENTING OBJECT RELATIONSHIPS IN A MULTIDIMENSIONAL SPACE

This application claims the benefit of U.S. Provisional Application No. 60/194,307, filed April 3, 2000.

Inventors: Dimitris K. Agrafiotis
Dmitrii N. Rassokhin
Victor S. Lobanov
F. Raymond Salemme

CROSS-REFERENCE TO RELATED APPLICATIONS

The following applications of common assignee are related to the present application, and are herein incorporated by reference in their entireties:

"System, Method and Computer Program Product for Representing Object Relationships in a Multidimensional Space," serial number (To be assigned, attorney ref. 1503.0870001), filed March 22, 2001;

"Method, System and Computer Program Product for Nonlinear Mapping of Multidimensional Data," serial number 09/303,671, filed May 3, 1999;

"System, Method and Computer Program Product for Representing Proximity Data in a Multidimensional Space," serial number 09/073,845, filed May 7, 1998;

"Method, System, and Computer Program Product for Representing Similarity/Dissimilarity Between Chemical Compounds," serial number 08/963,872, filed November 4, 1997;

"System, Method and Computer Program Product for the Visualization and Interactive Processing and Analysis of Chemical Data," serial number 08/963,872, filed November 4, 1997; and

"Stochastic Algorithms for Maximizing Molecular Diversity," serial number 60/030,187, filed November 4, 1996.

BACKGROUND OF THE INVENTION

Field of the Invention

[0001] The present invention relates to information representation, information cartography and data mining. The present invention also relates to pattern analysis and representation, and, in particular, representation of object relationships in a multidimensional space.

Related Art

[0002] Similarity is one of the most ubiquitous concepts in science. It is used to analyze and categorize phenomena, rationalize behavior and function, and design new entities with desired or improved properties. It is employed in virtually all scientific and technical fields, and particularly in data mining and information retrieval. Similarity (or dissimilarity) is typically quantified in the form of a numerical index derived either through direct observation, or through the measurement of a set of characteristic attributes, which are subsequently combined in some form of similarity or distance measure. For large collections of objects, similarities are usually described in the form of a matrix that contains some or all of the pairwise relationships between the objects in the collection. Unfortunately, pairwise similarity matrices do not lend themselves for numerical processing and visual inspection. A common solution to this problem is to embed the objects into a low-dimensional Euclidean space in a way that preserves the original pairwise relationships as faithfully as possible. This approach, known as multidimensional scaling (MDS) (Torgeson, W. S., *Psychometrika 17*:401-419 (1952); Kruskal, J. B., *Phychometrika 29*:115-129 (1964)) or nonlinear mapping (NLM) (Sammon, J. W., *IEEE Trans. Comp. C18*:401-409 (1969)), converts the data points into a set of real-valued vectors that can subsequently be used for a variety of pattern recognition and classification tasks.

[0003] Multidimensional scaling originated in the field of mathematical psychology and has two primary applications: 1) reducing the dimensionality of high-dimensional data in a way that preserves the original relationships of the data objects, and 2) producing Cartesian coordinate vectors from data supplied directly in the form of similarities or proximities, so that they can be analyzed with conventional statistical and data mining techniques.

[0004] Given a set of k objects, a symmetric matrix, r_{ij}, of relationships between these objects, and a set of images on a m-dimensional display plane $\{y_i, i = 1, 2, .., k; y_i \in \Re^m\}$, the problem is to place y_i onto the plane in such a way that their Euclidean distances $d_{ij} = \|y_i - y_j\|$ approximate as closely as possible the corresponding values r_{ij}. The quality of the projection is determined using a sum-of-squares error function such as Kruskal's stress:

$$S = \sqrt{\frac{\sum_{i<j}(d_{ij} - r_{ij})^2}{\sum_{i<j} r_{ij}^2}} \qquad (1)$$

which is numerically minimized in order to find the optimal configuration. The actual embedding is carried out in an iterative fashion by: 1) generating an initial set of coordinates y_i, 2) computing the distances d_{ij}, 3) finding a new set of coordinates y_i using a steepest descent algorithm such as Kruskal's linear regression or Guttman's rank-image permutation, and 4) repeating steps 2 and 3 until the change in the stress function falls below some predefined threshold.

[0005] A particularly popular implementation is Sammon's nonlinear mapping algorithm (Sammon, J. W. IEEE Trans. Comp., 1969). This method uses a modified stress function:

$$E = \frac{\sum_{i<j}^{k} \frac{[r_{ij} - d_{ij}]^2}{r_{ij}}}{\sum_{i<j}^{k} r_{ij}} \qquad (2)$$

which is minimized using steepest descent. The initial coordinates, y_i, are determined at random or by some other projection technique such as principal component analysis, and are updated using Eq. 3:

$$y_{ij}(t+1) = y_{ij}(t) - \lambda \Delta_{ij}(t) \qquad (3)$$

where t is the iteration number and λ is the learning rate parameter, and

$$\Delta_{ij}(t) = \frac{\dfrac{\partial E(t)}{\partial y_{ij}(t)}}{\left|\dfrac{\partial^2 E(t)}{\partial y_{ij}(t)^2}\right|} \qquad (4)$$

[0006] There is a wide variety of MDS algorithms involving different error functions and optimization heuristics, which are reviewed in Schiffman, Reynolds and Young, *Introduction to Multidimensional Scaling*, Academic Press, New York (1981); Young and Hamer, *Multidimensional Scaling: History, Theory and Applications*, Erlbaum Associates, Inc., Hillsdale, NJ (1987); Cox and Cox, *Multidimensional Scaling*, Number 59 in *Monographs in Statistics and Applied Probability*, Chapman-Hall (1994), and Borg, I., Groenen, P., Modern Multidimensional Scaling, Springer-Verlag, New York, (1997). The contents of these publications are incorporated herein by reference in their entireties.

[0007] Unfortunately, the quadratic nature of the stress function (Eqs. 1 and 2, and their variants) make these algorithms impractical for large data sets containing more than a few hundred to a few thousand items. Several attempts have been devised to reduce the complexity of the task. Chang and Lee (Chang, C. L., and Lee, R. C. T., *IEEE Trans. Syst., Man, Cybern.*, 1973, *SMC-3*, 197-200) proposed a heuristic relaxation approach in which a subject of the original objects (the frame) are scaled using a Sammon-like methodology, and the remaining objects are then added to the map by adjusting their distances to the objects in the frame. An alternative approach proposed by Pykett (Pykett, C. E., *Electron. Lett.*, 1978, *14*, 799-800) is to partition the data into a set of disjoint clusters, and map only the cluster prototypes, i.e. the centroids of the pattern vectors in each class. In the resulting two-dimensional plots, the cluster prototypes are represented as circles whose radii are proportional to the spread in their respective classes. Lee, Slagle and Blum (Lee, R. C. Y., Slagle, J. R., and Blum, H., *IEEE Trans.*

Comput., 1977, *C-27*, 288-292) proposed a triangulation method which restricts attention to only a subset of the distances between the data samples. This method positions each pattern on the plane in a way that preserves its distances from the two nearest neighbors already mapped. An arbitrarily selected reference pattern may also be used to ensure that the resulting map is globally ordered. Biswas, Jain and Dubes (Biswas, G., Jain, A. K., and Dubes, R. C., *IEEE Trans. Pattern Anal. Machine Intell.*, 1981, *PAMI-3(6)*, 701-708) later proposed a hybrid approach which combined the ability of Sammon's algorithm to preserve global information with the efficiency of Lee's triangulation method. While the triangulation can be computed quickly compared to conventional MDS methods, it tries to preserve only a small fraction of relationships, and the projection may be difficult to interpret for large data sets.

[0008] The methods described above are iterative in nature, and do not provide an explicit mapping function that can be used to project new, unseen patterns in an efficient manner. The first attempt to encode a nonlinear mapping as an explicit function is due to Mao and Jain (Mao, J., and Jain, A.K., *IEEE Trans. Neural Networks* 6(2):296-317 (1995)). They proposed a 3-layer feed-forward neural network with n input and m output units, where n and m are the number of input and output dimensions, respectively. The system is trained using a special back-propagation rule that relies on errors that are functions of the inter-pattern distances. However, because only a single distance is examined during each iteration, these networks require a very large number of iterations and converge extremely slowly.

[0009] An alternative methodology is to employ Sammon's nonlinear mapping algorithm to project a small random sample of objects from a given population, and then "learn" the underlying nonlinear transform using a multilayer neural network trained with the standard error back-propagation algorithm or some other equivalent technique (see for example, Haykin, S. Neural Networks: A Comprehensive Foundation. Prentice-Hall, 1998). Once trained, the neural network can be used in a feed-forward manner to project

the remaining objects in the plurality of objects, as well as new, unseen objects. Thus, for a nonlinear projection from n to m dimensions, a standard 3-layer neural network with n input and m output units is used. Each n-dimensional object is presented to the input layer, and its coordinates on the m-dimensional nonlinear map are obtained by the respective units in the output layer (Pal, N. R. Eluri, V. K., *IEEE Trans. Neural Net.*, 1142-1154 (1998)).

[00010] The distinct advantage of this approach is that it captures the nonlinear mapping relationship in an explicit function, and allows the scaling of additional patterns as they become available, without the need to reconstruct the entire map. However, as it was originally proposed, the method can only be used for dimension reduction, and requires that the input patterns be supplied as real vectors.

[00011] Hence there is a need for a method that can efficiently process large data sets, e.g., data sets containing hundreds of thousands to millions of items, and can be used with a wide variety of pattern representations and/or similarity distance functions. Moreover, there is a need for a method that is incremental in nature, and allows the mapping of new samples as they become available, without the need to reconstruct an entire map.

SUMMARY OF THE INVENTION

[00012] The present invention provides a method, system, and computer program product for representing a set of objects in a multidimensional space given a set of pairwise relationships between some of these objects.

[00013] In an embodiment, a plurality of objects are selected for comparison. At least some of all possible pairs of objects from the selected plurality of objects are compared, and the resulting pairwise relationships are recorded in a database. These pairwise relationships are used to embed the selected plurality of objects into an m-dimensional Euclidean space in such a way that the proximities (distances) of the selected objects in the m-dimensional Euclidean

space approximate as closely as possible the corresponding pairwise relationships. Hereafter, we shall refer to the coordinates of objects in this *m*-dimensional Euclidean space as "output coordinates" or "output features". In addition, a set of *n* attributes are determined for each of the selected plurality of objects. Hereafter, we shall refer to these *n* attributes as "input coordinates" or "input features". Thus, each of the selected plurality of objects is associated with an *n*-dimensional vector of input features and an *m*-dimensional vector of output features. A supervised machine learning approach is then employed to determine a functional relationship between the *n*-dimensional input and *m*-dimensional output vectors, and that functional relationship is recorded. Hereafter, we shall refer to this functional relationship as "mapping function".

[00014] In an embodiment, a pairwise relationships database is created by having one or more humans, apparatuses or computer processes initially select a plurality of objects that are presented to one or more humans, apparatuses or computer processes for pairwise comparison. Hereafter, the one or more humans, apparatuses or computer processes to which the objects are presented will be referred to as "subjects". In addition, one or more experts, apparatuses or computer processes select a set of *n* attributes or input features used to characterize each of the selected plurality of objects. A computer program product is used to select pairs of objects from the selected plurality of objects, and present these pairs of objects to the subjects. The computer program product then receives data from the subjects about the relationships of the objects presented to them. The relationship data from the subjects is recorded in the database and is used to embed the selected plurality of objects into an *m*-dimensional Euclidean space in such a way that the proximities (distances) of the selected plurality objects in that *m*-dimensional space approximate their corresponding relationships to a significant extent.

[00015] In an embodiment, output features for additional objects that were not part of the selected plurality of objects may be determined by computing their input features and evaluating the mapping function determined from the original relationship data using the supervised machine learning approach.

[00016] In an embodiment, relationships between the selected plurality of objects or additional objects that were not part of the selected plurality of objects may be determined by embedding the objects into the *m*-dimensional space and measuring their corresponding distances in that *m*-dimensional space.

[00017] In an embodiment, the mapping function is determined using one or more artificial neural networks.

[00018] In an embodiment, the *n* input features associated with an object represent the relationships of that object to *n* other reference objects.

[00019] Further embodiments, features, and advantages of the present invention, as well as the structure and operation of the various embodiments of the present invention, are described in detail below with reference to the accompanying figures.

BRIEF DESCRIPTION OF THE DRAWINGS/FIGURES

[00020] FIG. 1 illustrates the process of comparing a selected object to each of a set of *n* reference objects and using the results of this comparison as a fixed-length (*n*-dimensional) real vector that can be used as an input to the supervised machine learning technique in the manner of the present invention.

[00021] FIG. 2 is a flowchart illustrating exemplary phases of the method of the invention.

[00022] FIG. 3 is a flowchart illustrating an exemplary training phase of the invention, according to an embodiment of the invention.

[00023] FIG. 4 is a flowchart illustrating the use of a fuzzy clustering methodology in the selection of reference patterns, according to an embodiment of the invention.

[00024] FIG. 5 illustrates the concept of Voronoi Cells, as used in an embodiment of the invention.

[00025] FIG. 6 is a flowchart illustrating the projection of input patterns, according to an embodiment of the invention.

[00026] FIG. 7 illustrates the operation of local neural networks, according to an embodiment of the invention.

[00027] FIG. 8 is a flowchart illustrating an exemplary training phase of the invention, according to an alternative embodiment.

[00028] FIG. 9 is a flowchart illustrating the projection of input patterns, according to an alternative embodiment of the invention.

[00029] FIG. 10 illustrates the operation of global and local neural networks, according to an alternative embodiment of the invention.

[00030] FIG. 11 illustrates an exemplary computing environment within which the invention can operate.

[00031] FIG. 12 illustrates an example computer network environment of the present invention.

[00032]

DETAILED DESCRIPTION OF THE INVENTION

[00033] Preferred embodiments of the present invention are now described with references to the figures, where like reference numbers indicate identical or functionally similar elements. Also in the figures, the left most digit(s) of each reference number corresponds to the figure in which the reference number is first used. While specific configurations and arrangements are discussed, it should be understood that this is done for illustrative purposes only. One skilled in the relevant art will recognize that other configurations and arrangements can be used without departing from the spirit and scope of the invention. It will also be apparent to one skilled in the relevant art that this invention can also be employed in a variety of other devices and applications.

Nomenclature

[00034] The method of object representation, as described herein, can find applications in a variety of fields. "Objects" include, for example, items found

in nature; chemical compounds; processes; machines; compositions of matter; articles of manufacture; electrical devices; mechanical devices; financial data; financial instruments; financial trends; financial related traits and characteristics; marketing data; consumer profiles; census information; voting demographics; software products; human traits and characteristics; scientific properties, traits, and characteristics; and other tangible or intangible items that can be characterized by numerical measurement.

[00035] "Pairwise relationships" or "pairwise relationship measurements" provide a numerical measure of the relationship between two objects. This relationship may represent the similarity or dissimilarity between two objects, or some other form of association. In one embodiment, pairwise relationships are determined using a function that takes as input numerical attributes that characterize two objects. Such functions can take a variety of forms including, but not limited to, the Minkowski metrics, Hamming distance, Tanimoto (or Jacard) coefficient, Dice coefficient, and many others.

[00036] These pairwise relationships are used to embed the selected plurality of objects into an *m*-dimensional Euclidean space in such a way that the proximities (distances) of the selected objects in the *m*-dimensional Euclidean space approximate as closely as possible the corresponding pairwise relationships. This process is referred to as "mapping" and the coordinates of the objects in the *m*-dimensional Euclidean space as "output coordinates" or "output features". In addition, a set of *n* attributes are determined for each of the selected plurality of objects. These attributes are referred to as "input coordinates" or "input features". The functional relationship between the *n*-dimensional input and *m*-dimensional output vectors determined by the supervised machine learning approach is referred to as "mapping function".

Overview

[00037] The present invention uses supervised machine learning techniques to map members of a set of objects in a multidimensional space in a manner that

preserves to a significant extent known relationships between some of these objects. In particular, this specification describes a system, method and computer program product for embedding a set of objects into an *m*-dimensional Cartesian coordinate space, in such a way that the proximities of the objects in that *m*-dimensional space approximate the corresponding pairwise relationships. In addition, this specification describes a system, method and computer program product for embedding additional objects that are not part of the selected plurality of objects into that *m*-dimensional space. Without limitation but by way of illustration only, objects include, for example, items found in nature; chemical compounds; processes; machines; compositions of matter; articles of manufacture; electrical devices; mechanical devices; financial data; financial instruments; financial trends; financial related traits and characteristics; marketing data; consumer profiles; census information; voting demographics; software products; human traits and characteristics; scientific properties, traits, and characteristics; and other tangible or intangible items that can be characterized by fields of data.

[0001] In an embodiment, the present invention involves four steps: 1) pairwise relationship data generation, in which pairwise relationships between objects in a selected plurality of objects are generated and recorded in a database, 2) nonlinear mapping, in which the selected plurality of objects are embedded into an *m*-dimensional space in such a way that the proximities of the objects in that *m*-dimensional space approximate the original pairwise relationships to a significant extent, and 3) machine learning, in which a set of *n* attributes are measured or computed for each of the objects, and a mapping function that encodes the relationship between these *n* attributes and the coordinates of the objects in the *m*-dimensional space is determined and recorded. Optionally, the same *n* attributes may be measured or computed for a new set of objects, and the mapping function may be evaluated to generate coordinates for this new set of objects in the *m*-dimensional space. Optionally, the coordinates of objects on the *m*-dimensional space may be used to determine their pairwise relationships.

[00038] In the following discussion, artificial neural networks are used as an exemplary embodiment of a supervised machine learning technique. However, it should be understood that they are used by way of example, and not limitation. It will be apparent to persons skilled in the relevant art that any other supervised machine learning technique capable of function approximation (e.g. partial least squares, etc) can be employed.

Pairwise Relationship Data Generation

[00039] According to the method of the invention, at least some of all possible pairs of objects (patterns) from a selected plurality of objects are compared, and the resulting pairwise relationships are recorded in a database. As would be apparent to one skilled in the relevant art given the discussion herein, there are a number of approaches that can be taken in accordance with the method of the invention to select objects to be compared.

[00040] When applying the method of the invention to the field of molecular similarity, for example, one approach for selecting objects (compounds) is to judiciously select a subset of diverse objects (compounds) that would serve to define a reasonable compound space for similarity/dissimilarity analysis. In an embodiment, a subset of about 100-1000 diverse compounds can be selected for pairwise comparison.

[00041] As would be apparent to one skilled in the relevant art given the discussion herein, objects selected in the above manner map from an *n*-dimensional vector space to an *m*-dimensional vector space. Objects selected in this manner and mapped to an *m*-dimensional vector space can serve as markers that define the properties associated with specific areas or regions of the m-dimensional vector space. New objects, which map to a specific area of the *m*-dimensional vector space, are likely to have properties similar to the selected compounds that map to the same specific area of the *m*-dimensional

vector space. Thus, objects that map to the same specific area of the *m*-dimensional vector space are likely to be similar or related. Objects that map to different areas of the *m*-dimensional vector space are likely to be dissimilar or unrelated. The Euclidean distances between mapped compounds is a measure of their similarity/dissimilarity or relatedness.

[00042] A second approach for selecting objects for comparison is to select (e.g., randomly) a subset of objects from a database of objects of particular interest and have one or more humans, apparatuses or computer processes perform a pairwise comparison of the objects. For example, one or more humans, apparatuses or computer processes can present pairs of the selected objects to one or more humans, apparatuses or computer processes (subjects) for pairwise comparison, and the results of these comparisons by the subjects can be recorded. As described below, a computer program product can used to select pairs of objects from the selected plurality of objects, and present these pairs of objects to the subjects. The computer program product then receives data from the subjects about the relationships of the objects presented to them. The relationship data from the subjects is recorded in the database and is used to embed the selected plurality of objects into an *m*-dimensional Euclidean space, as described herein, in such a way that the proximities (distances) of the selected plurality objects in that *m*-dimensional space approximate their corresponding relationships to a significant extent.

[00043] Applying the second approach above to the field of molecular similarity, for example, compounds can be selected either at random, systematically, or semi-systematically from a computer database and presented to one or more chemists for pairwise comparison. The number of compounds selected should be statistically chosen so that the selected compounds are likely to be representative of the compounds in the database of interest. Thus, depending on the size of the database and the method used to select compounds, the number of compounds selected for comparison by the one or more chemists will vary. For example, in an embodiment about 100 to

1000 compounds might be selected for comparison by the one or more chemists.

[00044] This second selection approach can be used, for example, in the field of molecular similarity to mine a database of compounds and identify compounds similar to compounds having known therapeutic, agricultural or other commercial value. As described herein, the compounds selected from the database can be multidimensionally scaled to an *m*-dimensional vector space and used to determine one or more nonlinear mapping functions. These mapping functions can then be used to map other compounds in the same or a different database to the *m*-dimensional vector space in order to determine which compounds in the database may be commercially valuable. Compounds having known therapeutic, agricultural or other commercial value can be selected and mapped to the *m*-dimensional vector space to identify particular areas or regions of importance. New compounds which map to the same area or region of the *m*-dimensional vector space as the compounds having known commercial or therapeutic value are likely to be similar to the compounds having the known commercial therapeutic, agricultural or other commercial value.

[00045] As described below, in an embodiment of the invention pairwise relationship data about objects is generated and recorded in a database using a computer network (e.g., the INTERNET). This aspect of the invention is illustrated herein using an example embodiment of the invention from the field of molecular similarity.

[00046] In an embodiment of the present invention, similarity data about compounds of interest is received via a computer network from a plurality of scientist (e.g., chemists). A computer program product running on a first computer is used to select pairs of objects (compounds) from a database, and these selected objects (compounds) are then sent from the first computer via a computer network to scientists at a remote location. At the remote location, the selected pairs of objects (compounds) are presented to the scientists, for example, by a second computer, as compounds for pairwise

similarity comparison. The scientists evaluate the similarity of the pairs of compounds presented to them and send data about the compounds via the computer network to the first computer.

[00047] In an embodiment, the data received from the scientists is stored in a database for subsequent retrieval. The pairwise similarity data from the scientists is used as described herein to develop a mapping function according to the invention. The mapping function can be derived using the similarity data from the scientists. As described herein, once determined, a mapping function according to the invention can be used to map additional compounds into the *m*-dimensional space, including compounds not presented to the scientists for similarity comparison, without the need for additional input from a scientist.

[00048] In an embodiment, the method of the invention comprises the steps of randomly selecting two compounds from the plurality of compounds selected for pairwise comparison; presenting the two compounds selected to a scientist; and receiving data from the scientist about the similarity of the compounds. These steps are repeated for additional pairs of compounds for as long as the scientist is willing to evaluate pairs of compounds. The selection of compounds in this embodiment can be performed by a computer program product running on an unattended computer. For example, in an embodiment, a computer program product running on a server receives similarity data from a plurality of scientists via the INTERNET.

[00049] The present invention is not limited to using a computer and a computer program product to receive similarity data. As would be apparent to persons skilled in the relevant art given the discussion herein, there are many other ways to receive similarity data about objects (e.g., a plurality of compounds). For example, a survey could be sent to a scientist by mail, which would ask the scientist to evaluate the similarity of compounds. Alternatively, a person could call a scientist using a telephone and ask the scientist questions about the compounds. Thus, it should be understood that the ways for

receiving data described herein are presented by way of example only, and not limitation.

[00050] In an embodiment, personal and/or background data about the subjects providing similarity data is collected and used to develop tailored mapping functions according to the invention. For example, personal and/or background data about the scientists providing the similarity data for compounds as described herein is received from the scientists. This data can include, for example, data about the scientist's education, training, work experiences, research interests, et cetera. This data can then be used to further analysis the similarity data received from the scientists (e.g, where data is received from a plurality of scientists). Personal and/or background data can be used, for example, to determine how a particular group of scientists would classify the similarity of certain compounds of interest compared to another groups of scientists. The personal and/or background data can be used, for example, to determine how chemists from a first field of chemistry would classify the similarity of a set of compounds compared to chemists from a second field of chemistry. Other determinations about how a first group of scientists would classify the similarity of a set of compounds compared to a second group of scientists can also be made, as would be apparent to a person skilled in the relevant art given the discussion herein. Thus, it should be understood that the ways for using the personal and/or background data described herein are presented by way of example only, and not limitation. Furthermore, this feature of the invention is not limited to the field of molecular similarity. As would be apparent to a person skilled in the relevant art given the discussion herein, this feature of the invention can be applied to any other field.

[00051] When using the method of the invention, it is expected that information about some selected objects might be unknown or unavailable. As described herein, the present invention can be implemented using data that is incomplete, noisy, and/or corrupt. To illustrate this feature of the invention, assume that the data in Table 1, below, was receive according to the method of

the invention from five subjects (i.e., S1, S2, S3, S4, and S5). The data in Table 1 represents how the five subjects might rate the similarity of five objects (A-E) on a scale of zero to one, where zero represents absolute similarity and one represent absolute dissimilarity. As can be seen by looking at Table 1, data about the similarity of the objects is incomplete, noisy, and/or corrupt.

[00052] In an embodiment of the invention, the similarity data from the five subjects is combined using averaging to produce the results in Table 2. This averaged data is then used to create a pairwise relationship database according to the invention. As would be apparent to a person skilled in the relevant art given the discussion herein, methods other than averaging can be used to combine the similarity data received from the subjects.

	S1	S2	S3	S4	S5
AB	0.5	0.7	0.6		
AC	0.7	0.9	0.8	0.8	0.7
AD	0.6	0.5	0.6	0.7	
AE	0.4	0.4	0.4	0.5	0.5
BC		0.5	0.7	0.6	
BD	0.9	0.7	0.8	0.8	0.7
BE	0.3	0.4	0.5	0.4	0.5
CD			0.5	0.7	0.6
CE	0.4	0.5	0.5	0.4	0.4
DE					

TABLE 1

AB	AC	AD	AE	BC	BD	BE	CD	CE	DE
0.60	0.78	0.60	0.44	0.60	O.78	0.42	0.60	0.44	

TABLE 2

[00053] It should be understood that the approaches for generating pairwise relationship data and recording the data in a database described above have been presented by way of example only, and not limitation. Other approaches will be apparent to persons skilled in the relevant art given the discussion herein.

Nonlinear Mapping Using Subset Refinements

[00054] A nonlinear mapping algorithm that is well suited for large data sets is presented in U.S. Patent Application 09/303,671, filed May 3, 1999, titled, "Method, System and Computer Program Product for Nonlinear Mapping of Multidimensional Data", and U.S. Patent Application 09/073,845, filed May 7, 1998, titled, "Method, System and Computer Program Product for Representing Proximity Data in a Multidimensional Space". This approach is to use iterative refinement of coordinates based on partial or stochastic errors.

[00055] The method uses a self-organizing principle to iteratively refine an initial (random or partially ordered) configuration of objects by analyzing only a subset of objects and their associated relationships at a time. The relationship data may be complete or incomplete (i.e. some relationships between objects may not be known), exact or inexact (i.e. some or all relationships may be given in terms of allowed ranges or limits), symmetric or asymmetric (i.e. the relationship of object A to object B may not be the same as the relationship of B to A) and may contain systematic or stochastic errors.

[00056] The relationships between objects may be derived directly from observation, measurement, a priori knowledge, or intuition, or may be

determined directly or indirectly using any suitable technique for deriving such relationships.

[00057] The invention determines the coordinates of a plurality of objects on the *m*-dimensional nonlinear map by:

(1) placing the objects on the *m*-dimensional nonlinear map;

(2) selecting a subset of the objects, wherein the selected subset of objects includes associated relationships between objects in the selected subset;

(3) revising the coordinate(s) of one or more objects in the selected subset of objects on the *m*-dimensional nonlinear map based on the relationship(s) between some of these objects and their corresponding distance(s) on the nonlinear map;

(4) repeating steps (2) and (3) for additional subsets of objects from the plurality of objects.

[00058] In one embodiment, subsets of objects can be selected randomly, semi-randomly, systematically, partially systematically, etc. As subsets of objects are analyzed and their distances on the nonlinear map are revised, the set of objects tends to self-organize.

[00059] In a preferred embodiment, the invention iteratively analyzes a pair of objects at a time, that is, step (2) is carried out by selecting a pair of objects having an associated pairwise relationship. Pairs of objects can be selected randomly, semi-randomly, systematically, partially systematically, etc. This embodiment is described herein for illustrative purposes only and is not limiting. Various other embodiments are described herein.

[00060] In a preferred embodiment, the method starts with an initial configuration of points generated at random or by some other procedure such as principal component analysis. This initial configuration is then continuously refined by repeatedly selecting two objects, *i*, *j*, at random, and modifying their coordinates on the nonlinear map according to Eq. 5:

$$y_i(t+1) = f(t, y_i(t), y_j(t), r_{ij}) \qquad (5)$$

where t is the current iteration, $y_i(t)$ and $y_j(t)$ are the current coordinates of the i-th and j-th objects on the nonlinear map, $y_i(t+1)$ are the new coordinates of the i-th object on the nonlinear map, and r_{ij} is the relationship between the i-th and j-th objects. $f(.)$ in Eq. 5 above can assume any functional form. Ideally, this function should try to minimize the difference between the distance on the nonlinear map and the actual relationship between the i-th and j-th objects. For example, $f(.)$ may be given by Eq. 6:

$$y_i(t+1) = 0.5\lambda(t)\frac{r_{ij} - d_{ij}(t)}{d_{ij}(t)}(y_i(t) - y_j(t)) \qquad (6)$$

where t is the iteration number, $d_{ij} = \|y_i(t) - y_j(t)\|$, and $\lambda(t)$ is an adjustable parameter, referred to hereafter as the "learning rate". This process is repeated for a fixed number of cycles, or until some global error criterion is minimized within some prescribed tolerance. A large number of iterations are typically required to achieve statistical accuracy.

[00061] The method described above is generally reminiscent of the error back-propagation procedure for training artificial neural networks described in Werbos, Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences, PhD Thesis, Harvard University, Cambridge, MA (1974), and Rumelhart and McClelland, Eds., Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1, MIT Press, Cambridge, MA (1986), both of which are incorporated herein by reference in their entireties.

[00062] The learning rate $\lambda(t)$ in EQ. 6 plays a key role in ensuring convergence. If λ is too small, the coordinate updates are small, and convergence is slow. If, on the other hand, λ is too large, the rate of learning may be accelerated, but the nonlinear map may become unstable (i.e. oscillatory). Typically, λ ranges in the interval [0, 1] and may be fixed, or it may decrease monotonically during the refinement process. Moreover, λ may also be a function of i, j, r_{ij}, and/or d_{ij}, and can be used to apply different weights to certain objects, relationships, distances and/or relationship or distance pairs.

[00063] One of the main advantages of this approach is that it makes partial refinements possible. It is often sufficient that the pairwise similarities are represented only approximately to reveal the general structure and topology of the data. Unlike traditional MDS, this approach allows very fine control of the refinement process. Moreover, as the nonlinear map self-organizes, the pairwise refinements become cooperative, which partially alleviates the quadratic nature of the problem.

[00064] The embedding procedure described above does not guarantee convergence to the global minimum (i.e., the most faithful embedding in a least-squares sense). If so desired, the refinement process may be repeated a number of times from different starting configurations and/or random number seeds.

[00065] The general algorithm described above can also be applied when the pairwise similarity matrix is incomplete, i.e. when some of the pairwise relationships are unknown, uncertain or corrupt, or both of the above. It may also be applied when the pairwise relationships are asymmetric, and when some of the pairwise relationships have multiple measurements. Various other embodiments are described herein.

Encoding

[00066] This present invention uses the iterative nonlinear mapping algorithm described above to multidimensionally scale a small random sample of the population of objects, and then "learns" the underlying nonlinear transform using an artificial neural network or some other suitable supervised machine learning technique. In the simple case of dimension reduction, the neural network takes as input the n input features associated with the object(s) of interest, and produces as output the coordinates of the objects on the m-dimensional nonlinear map. The present invention is aimed at cases where the object representation that is used to determine the relationships between objects is not in the form of an n-dimensional real vector, and therefore cannot be used directly as input to the neural network. Consider for example, the

relationship (similarity) between two conformations of an organic molecule. A common measure of similarity is the root-mean-squared deviation of the atomic coordinates after a least-squares superposition of the two conformations. The object representation (atomic coordinates) can not be used directly as input to the neural network, and needs to be recast in the form of an n-dimensional real vector. Another example represents relationships (similarities) between long binary encoded objects, i.e. objects described by a long sequence of binary numbers, such as images, for example. Because of their high dimensionality, such long binary representations do not lend themselves as input to a neural network.

[00067] Thus, the present invention attempts to convert each object into a fixed-length (n-dimensional) real vector that can be used as input to the supervised machine learning technique. This can be carried out in a variety of ways. In a preferred embodiment, a set of n reference objects is identified, and every object in the plurality of objects is compared to these n reference objects. Thus, every object is associated with n relationships to a fixed set of reference objects. These n relationships, along with the m output coordinates on the nonlinear map, are then used as input to a supervised machine learning technique, which extracts the mapping function after a process of training. The process is illustrated in FIG. 1.

[00068] In another embodiment, a set of n object attributes are determined for each object. Ideally, these attributes should be related to the method used to determine the relationships between objects. For example, in the case of long binary representations, the n input features may represent the n principal components of the binary feature matrix that account for most of the variance in the data set. In general, the more relevant the n input features are to the problem of interest, the better the supervised machine learning technique is able to approximate the mapping function.

[00069]

Nonlinear Mapping Networks – Algorithm I

[00070] In an exemplary embodiment, a simple 3-layer network with *n* input and *m* output units can be employed to determine the mapping function. The network is trained to reproduce the input/output features of the objects produced by the iterative nonlinear mapping algorithm, and thus encodes the mapping in its synaptic parameters in a compact, analytical manner. Once trained, the neural network can be used in a feed-forward fashion to project the remaining members of the input set, as well as new, unseen objects.

[00071] In an embodiment, the projection can be carried out using a multiplicity of neural networks, each of which is trained independently and specializes in the prediction of a subset of the *m* output features. For example, the system may involve *m* independent neural networks each of which specializes in the prediction of a single output feature.

[00072] In an embodiment, a feature selection algorithm may also be employed in order to reduce the number of inputs supplied to the neural network(s). A feature selection algorithm identifies a subset of the original *n* input features that are sufficient for training. Many feature selection algorithms can be employed, ranging from greedy approaches, to Monte-Carlo search, simulated annealing and genetic algorithms. It will be apparent to a person skilled in the relevant art how to implement a feature selection algorithm in the context of the present invention.

[00073] The method of the invention is illustrated generally in FIG. 2. The method begins at step 205. In step 210, the training of a neural network takes place, where the training is based on the results (i.e., the inputs and outputs) of the iterative algorithm. In step 215, points in R^n are projected into R^m by a feed-forward pass through the trained neural network The process concludes with step 220.

Local Nonlinear Mapping Networks – Algorithm II

[00074] The embodiment of the invention described in this section represents a variation of the above algorithm. This approach is based on local learning. Instead of using a single "global" network to perform the nonlinear mapping across the entire input data space R^n, this embodiment partitions the space into a set of Voronoi polyhedra, and uses a separate "local" network to project the patterns in each partition. Given a set of reference points $P = \{\mathbf{p}_1, \mathbf{p}_2, \ldots\}$ in R^n, a Voronoi polyhedron (or Voronoi cell), $v(\mathbf{p})$, is a convex polytope associated with each reference point p which contains all the points in R^n that are closer to $\mathbf{p}$ than any other point in P:

$$v(\mathbf{p}) = \{\mathbf{x} \in R^n \mid d(\mathbf{x},\mathbf{p}) \leq d(\mathbf{x},\mathbf{q}) \quad \forall \mathbf{p},\mathbf{q} \in P, \mathbf{p} \neq \mathbf{q}\} \tag{11}$$

where $d()$ is a distance function. In an embodiment of the invention, $d()$ is the Euclidean distance function. Voronoi cells partition the input data space R^n into local regions "centered" at the reference points P, also referred to as centroids. Hereafter, the local networks associated with each Voronoi cell are said to be centered at the points P, and the distance of a point in R^n from a local network will refer to the distance of that point from the network's center.

[00075] The training phase involves the following general steps: a training set is extracted from the set of input patterns and mapped using the iterative nonlinear mapping algorithm described above. A set of reference points in the input space R^n is then selected, and the objects comprising the training set are partitioned into disjoint sets containing the patterns falling within the respective Voronoi cells. Patterns that lie on the sides and vertices of the Voronoi cells (i.e. are equidistant to two or more points in P), are arbitrarily assigned to one of the cells. A local network is then assigned to each cell, and is trained to reproduce the input/output mapping of the input patterns in that cell. While the direct nonlinear map is obtained globally, the networks are trained locally using only the input patterns within their respective Voronoi partitions. Again, simple 3-layer neural network with n

input and m output units can be employed, where n and m are the dimensionalities of the input and output spaces, respectively.

[00076] The training phase of the method of the invention therefore involves the following steps as illustrated in FIG. 3. The training phase begins at step 305. In step 310, a random set of points $\{\mathbf{x}_i, i=1,2...,k; \mathbf{x}_i \in R^n\}$ is extracted from the set of input patterns. In step 315, the points $\mathbf{x}_i$ are mapped from R^n to R^m using the iterative nonlinear mapping algorithm described above ($\mathbf{x}_i \rightarrow \mathbf{y}_i$, i = 1,2,...,k, $\mathbf{x}_i \in R^n$, $\mathbf{y}_i \in R^m$). This mapping serves to define a training set T of ordered pairs $(\mathbf{x}_i, \mathbf{y}_i)$, $T = \{(\mathbf{x}_i, \mathbf{y}_i), i = 1,2,...,k\}$.

[00077] In step 320, a set of reference points $P = \{\mathbf{c}_i, i = 1,2,...c; \mathbf{c}_i \in R^n\}$ is determined. In an embodiment of the invention, the reference points $\mathbf{c}_i$ are determined using a clustering algorithm described in greater detail below. In step 325, the training set T is partitioned into c disjoint clusters based on the distance of each pattern $\mathbf{x}_i$ from each reference pattern. The set of disjoint clusters is denoted $\{C_j = \{(\mathbf{x}_i, \mathbf{y}_i): d(\mathbf{x}_i, \mathbf{c}_j) \leq d(\mathbf{x}_i, \mathbf{c}_k)\}$ for all $k \neq j$; j = 1,2,...,c; i = 1,2,...,k}}. In step 330, c independent local networks $\{Net_i^L, i = 1,2,...,c\}$ are trained with the respective training subsets C_i derived in step 325. The training phase concludes with step 335.

[00078] Clearly, an important choice to be made concerns the partitioning. In general, the reference points $\mathbf{c}_i$ (determined in step 320) should be well distributed and should produce balanced partitions that contain a comparable number of training patterns. This is necessary in order to avoid the creation of poorly optimized networks due to an insufficient number of training cases. In one embodiment, described here, the reference points $\mathbf{c}_i$ can be determined using a clustering methodology.

[00079] Once determined, the reference points $\mathbf{c}_i$ are used to partition the input data set into a set of Voronoi cells. Such cells are illustrated in FIG. 5. A set 500 is shown partitioned into Voronoi cells, such as cells 505A through 505C. The Voronoi cells include refence points 510A through 510C respectively.

[00080] Once all the local networks are trained, additional patterns from the input set of patterns can be mapped into R^m as illustrated in FIG.6. The process begins with step 605. In step 610, the distance of the input pattern $\mathbf{x}$ to each reference point in $\{\mathbf{c}_i, i = 1,2,...c; \mathbf{c}_i \in R^n\}$ is determined. In step 615, the point $\mathbf{c}_j$ that is nearest to the input pattern $\mathbf{x}$ is identified. In step 620, the pattern $\mathbf{x}$ is mapped to a point $\mathbf{y}$ in R^m, $\mathbf{x} \rightarrow \mathbf{y}$, $\mathbf{x} \in R^n$, $\mathbf{y} \in R^m$ using the local neural network Net_j^L associated with the reference point $\mathbf{c}_j$ identified in step 615. The process concludes with step 625.

[00081] Note that new patterns in R^n that not in the original input set can also be projected into R^m in the manner shown in FIG. 6. Once the system is trained, new paterns in R^n are mapped by identifying the nearest local network and using that network in a feed-forward manner to perform the projection. An embodiment of a system that does this is illustrated in FIG. 7. The input for the system is a pattern 705 in R^n. This point is defined by its n attributes, $(x_1, x_2, ...x_n)$. The system includes a dispatcher module 710, which compares the distance of the input point to the network centers (i.e., the reference points), and forwards the input point to one of the available local neural networks 701, 702, or 703. Specifically, the input pattern is sent to the local neural network associated with the reference point nearest to the input pattern. The chosen network then performs the final projection, resulting in an output point in R^m.

Local Nonlinear Mapping Networks – Algorithm III

[00082] The ability of a single network to reproduce the general structure of the nonlinear map suggests an alternative embodiment to overcome some of the complexities of clustering in higher dimensions. Conceptually, the alternative embodiment differs from Algorithm II in the way it partitions the data space. In contrast to the previous method, this process partitions the output space, and clusters the training patterns based on their proximity on the m-dimensional nonlinear map rather than their proximity in

the *n*-dimensional input space. For the training set, the assignment to a partition is straightforward. The images of the points in the training set on the nonlinear map are derived directly from the iterative nonlinear mapping algorithm described above. For new points that are not part of the training set, the assignment is based on approximate positions derived from a global neural network trained with the entire training set, like the one of Algorithm II described above. A. The general flow of the algorithm is similar to the one described in the preceding section.

[00083] The training phase for this embodiment is illustrated in FIG. 8. The method begins at step 805. In step 810, a random set of patterns $\{\mathbf{x}_i, i = 1,2,\ldots,k; x_i \in R^n\}$ is extracted from the input data set. In step 815, the patterns $\mathbf{x}_i$ are mapped from R^n to R^m using the iterative nonlinear mapping algorithm described in section I B ($\mathbf{x}_i \rightarrow \mathbf{y}_i$, $i = 1,2,\ldots,k$, $\mathbf{x}_i \in R^n$, $\mathbf{y}_i \in R^m$). This mapping serves to define a training set T of ordered pairs $(\mathbf{x}_i, \mathbf{y}_i)$, $T = \{(\mathbf{x}_i, \mathbf{y}_i), i = 1,2,\ldots,k\}$.

[00084] In step 820, the points $\{\mathbf{y}_i, i = 1,2,\ldots,k, \mathbf{y}_i \in R^m\}$ are clustered into c clusters associated with c points in R^m, $\{\mathbf{c}_i, i = 1,2,\ldots c; \mathbf{c}_i \in R^m\}$. In the illustrated embodiment, fuzzy clusters are formed in this step using the FCM algorithm of FIG. 4. In step 825, the training set T is partitioned into c disjoint clusters C_j based on the distance of the images $\mathbf{y}_i$ from the cluster prototypes, $\{C_j = \{(\mathbf{x}_i, \mathbf{y}_i): d(\mathbf{y}_i, \mathbf{c}_j) \leq d(\mathbf{y}_i, \mathbf{c}_k)$ for all $k \neq j$; $j = 1,2,\ldots,c$; $i = 1,2,\ldots,k\}\}$. In step 830, c independent local neural networks $\{Net_i^L, i = 1,2,\ldots,c\}$ are trained with the respective clusters C_j derived in step 825. In step 835, a global network Net^G is trained with the entire training set T. The process concludes with step 840.

[00085] Once all the networks are trained, remaining input patterns from the input data set and any new patterns in R^n are projected using a tandem approach. An embodiment of this is illustrated in FIG. 9. The projection process begins at step 905. In step 910, each input pattern $\mathbf{x}$ to be projected into R^m is mapped, $\mathbf{x} \rightarrow \mathbf{y}'$, $\mathbf{x} \in R^n$, $\mathbf{y}' \in R^m$, using the global network Net^G derived in step 835.

[00086] In step 915, the distance from $\mathbf{y}'$ to each reference point $\mathbf{c}_i$ in $\{\mathbf{c}_i, i = 1,2,\ldots c; c_i \in R^m\}$ is determined. In step 920, the point $\mathbf{c}_j$ closest to $\mathbf{y}'$ is determined. In step 925, $\mathbf{x}$ is mapped into R^m, $\mathbf{x} \rightarrow \mathbf{y}$, using the local neural network associated with $\mathbf{c}_j$, Net_j^L. The process ends with step 930.

A system for performing the overall mapping $\mathbf{x} \rightarrow \mathbf{y}$ is shown in FIG. 10. First, an input pattern 1005 $\in R^n$ is projected by the global network 1010, Net^G, to obtain point 1012 $(\mathbf{y}') \in R^m$. Point $\mathbf{y}'$ can be viewed as having approximate coordinates on the nonlinear map. These coordinates are used to identify the nearest local network 1021 (Net_j^L) from among the possible local neural networks 1021 through 1023, based on the proximity of $\mathbf{y}'$ to each $\mathbf{c}_i$. Input point 1005 is projected once again, this time by the nearest local network 1021 to produce the final image 1030 on the display map.

Environment

[0100] FIG. 11 shows an example computer system 1100 that supports implementation of the present invention. The present invention may be implemented using hardware, software, firmware, or a combination thereof. It may be implemented in a computer system or other processing system. The computer system 1100 includes one or more processors, such as processor 1104. The processor 1104 is connected to a communication infrastructure 1106 (e.g., a bus or network). Various software embodiments can be described in terms of this exemplary computer system. After reading this description, it will become apparent to a person skilled in the relevant art how to implement the invention using other computer systems and/or computer architectures.

[0101] Computer system 1100 also includes a main memory 1108, preferably random access memory (RAM), and may also include a secondary memory 1110. The secondary memory 1110 may include, for example, a hard disk drive 1012 and/or a removable storage drive 1114, representing a floppy disk

drive, a magnetic tape drive, an optical disk drive, etc. The removable storage drive 1114 reads from and/or writes to a removable storage unit 1118 in a well known manner. Removable storage unit 1118 represents a floppy disk, magnetic tape, optical disk, etc. As will be appreciated, the removable storage unit 1118 includes a computer usable storage medium having stored therein computer software and/or data. In an embodiment of the invention, removable storage unit 1118 can contain input data to be projected.

[0102] Secondary memory 1110 can also include other similar means for allowing computer programs or input data to be loaded into computer system 1100. Such means may include, for example, a removable storage unit 1122 and an interface 1120. Examples of such may include a program cartridge and cartridge interface (such as that found in video game devices), a removable memory chip (such as an EPROM, or PROM) and associated socket, and other removable storage units 1122 and interfaces 1120 which allow software and data to be transferred from the removable storage unit 1122 to computer system 1100.

[0103] Computer system 1100 may also include a communications interface 1124. Communications interface 1124 allows software and data to be transferred between computer system 1100 and external devices. Examples of communications interface 1124 may include a modem, a network interface (such as an Ethernet card), a communications port, a PCMCIA slot and card, etc. Software and data transferred via communications interface 1124 are in the form of signals 1128 which may be electronic, electromagnetic, optical or other signals capable of being received by communications interface 1124. These signals 1128 are provided to communications interface 1124 via a communications path (i.e., channel) 1126. This channel 1126 carries signals 1128 and may be implemented using wire or cable, fiber optics, a phone line, a cellular phone link, an RF link and other communications channels. In an embodiment of the invention, signals 1128 can include input data to be projected.

[0104] Computer programs (also called computer control logic) are stored in main memory 1108 and/or secondary memory 1110. Computer programs may also be received via communications interface 1124. Such computer programs, when executed, enable the computer system 1100 to perform the features of the present invention as discussed herein. In particular, the computer programs, when executed, enable the processor 1104 to perform the features of the present invention. Accordingly, such computer programs represent controllers of the computer system 1100.

Computer Network Environment

[0105] FIG. 12 is a block diagram illustrating an exemplary computer network or network system 1200. Network system 1200 is used to receive pairwise relationship data in accordance with an embodiment of the present invention. Network server 1201 is shown having multiple central processor units (CPU) 1202. In other embodiments, network server 1201 can have only one CPU 1202. Computers 1230 are shown having only a single CPU 1232, but they can have more than one CPU 1232. Portions of the present invention are comprised of computer-readable and computer executable instructions which reside in computer-usable media of network server 1201 and/or computers 1230.

[0106] Network server 1201 has multiple central processor units (CPU) 1202. CPUs 1202 are electrically connected to a communication infrastructure 1203. Also connected to communications infrastructure 1203 are a main memory unit 1208, a secondary memory unit 1210, a graphics subsystem 1212, and a communications interface unit 1214. Each CPU 1202 may have a cache memory unit 1204. Network server 1201 is typically optimized for managing network communications, storing files, and/or processing database queries.

[0107] Computers 1230 are depicted as having only a single CPU 1202. Computers 1230 may have more than one CPU, however. CPU 1202 is

electrically connected to a communication infrastructure 1233. Also connected to communications infrastructure 1233 are a communications interface unit 1234, an input device 1236, and a display 1238.

[0108] The user inputs commands to computer 1230 using input device 1236. The commands are then either processed by CPU 1232 or transferred to another CPU for processing via communications interface 1234, communications link 1222, and network 1220.

[0109] As described herein, in an embodiment, a computer program product running on network server 1201 is used to select pairs of objects (patterns) from a database. These selected objects are then sent from network server 1201 via network 1220 or communications link 1222 to one or more subjects at computers 1230. At computers 1230, the selected pairs of objects are presented to the subjects on displays 1238 as objects for pairwise comparison. The subjects evaluate the similarity of the pairs of objects presented to them and enter data using input devices 1236. The data input by the subjects is then send via network 1220 or communications link 1222 to network server 1201. In an embodiment, the data received from the subjects is stored in a database in network server 1201 for subsequent retrieval and use in accordance with the invention.

[0110]

Conclusion

[0111] While various embodiments of the present invention have been described above, it should be understood that they have been presented by way of example, and not limitation. It will be apparent to persons skilled in the relevant art that various changes in detail can be made therein without departing from the spirit and scope of the invention. Thus the present invention should not be limited by any of the above-described exemplary

embodiments, but should be defined only in accordance with the following claims and their equivalents.